Contrasting skin color bias in skin lesion segmentation using UNet and SCSE - Group 5

Luca Barda
4746753
M.Sc. Computer Science Engineering
luca.barda@mail.polimi.it

Damiano Baschiera
4746749
M.Sc. Mathematical Engineering
damiano.baschiera@mail.polimi.it

Abstract

A major challenge in dermatological image analysis is the underrepresentation of dark-skinned patients in training datasets. We develop and evaluate data augmentation techniques for skin lesion segmentation across diverse skin tones, while also evaluating UNet architectures enhanced with Squeeze-and-Excitation (SCSE) and channel Squeeze-and-Excitation (cSE) attention mechanisms. We apply region-specific augmentations to generate synthetic image variations able to enhance the generalizability and performance of the models.

We assess the impact of augmentation strategies and architectural modifications on segmentation accuracy across different skin types. The UNet enriched with cSE blocks trained on the largest augmented set achieves the best performance. Although our framework is not yet clinically applicable, it provides a promising foundation for real-world deployment. Future work should focus on refining region-specific augmentations and expanding dataset diversity to enhance its clinical utility.

1. Introduction

Skin cancer is the most common cancer in the United States [8], with current estimates suggesting that one in five Americans will develop it during their lifetime [19]. Early detection significantly improves treatment outcomes, as most cases are highly treatable when identified promptly, especially for melanoma [9]. Accurate segmentation of skin lesions is a crucial step in computer-aided diagnosis systems, as it facilitates precise analysis of lesion characteristics, thereby aiding in early detection and treatment planning [11]. Deep neural networks have been widely applied to skin lesion segmentation with demonstrated effectiveness [14]. However, these networks require substantial high-quality training data to perform optimally. A significant limitation in current skin lesion datasets is the underrepresentation of dark-skinned patients. For instance, the widely-used HAM10000 dataset contains less than 5% of images representing dark skin tones [15]. This imbalance raises concerns about diagnostic equity and algorithm performance across diverse populations.

2. Related Works

The application of UNet architectures in binary segmentation of skin lesions is well-established in the literature [14, 1, 2, 6, 3, 4, 12]. Numerous studies have demonstrated the efficacy of UNets and ResNets for this task. For instance, Mirikharaji *et al*. [14] conducted a comprehensive survey on deep learning for skin lesion segmentation, analyzing 177 research papers and highlighting the prevalence of UNet-based models in achieving state-of-the-art results. Our project builds upon the UNet architecture implemented by Ashraf *et al*. [1], who developed a melanoma detection framework utilizing a modified UNet model, achieving high segmentation accuracy. However, recent research has identified significant biases in segmenting darker skin tones when employing conventional UNet techniques. Benčević *et al*. [2] conducted a thorough analysis using various image labeling and skin tone classification methods, concluding that skin lesion segmentation models available through 2023 performed inadequately when applied to diverse patient populations in real-world settings. Their study revealed a significant correlation between segmentation performance and skin color, with models consistently underperforming on darker skin tones. Additionally, Daneshjou *et al*. [6] highlighted disparities in dermatology AI performance on diverse clinical image sets, emphasizing the need for more inclusive datasets and model training to mitigate biases. To address these biases, researchers have explored multiple approaches. Bissoto *et al*. [3] proposed black-box techniques to construct unbiased datasets by balancing skin tone representation, aiming to reduce model biases without al-

Dataset	# Images	# Masks	Duplicates Removed
ISIC 2016 (train+test)	1279	1279	–
ISIC 2017 (train+val+test)	2750	2750	–
ISIC-Merged	2975	2975	1054

Table 1: Image and segmentation mask counts in ISIC 2016 [7] and ISIC 2017 [5] (all splits) , and in the final merged dataset.

tering the model architecture. In another study, Bissoto *et al*. [4] implemented Generative Adversarial Networks (GANs) to generate synthetic skin lesion images with diverse skin tones, enhancing the training data diversity and improving model generalization. Similarly, Mikolajczyk *et al*. [12] utilized GAN-based methods to bias the training process towards underrepresented skin tones, thereby mitigating segmentation biases. In contrast to these complex methodologies, our approach investigates the efficacy of simple image transformations based on lesion masks, utilizing existing segmentation masks from ground truth data. Specifically, we aim to assess whether such targeted augmentation techniques can enhance segmentation performance compared to basic data augmentation methods, considering their simplicity and computational efficiency.

3. Approach

3.1. Data

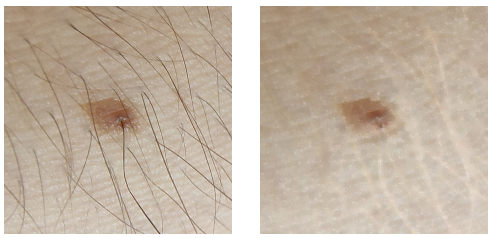
In this project, we use a joint dataset constructed from all available data splits of the International Skin Imaging Collaboration (ISIC) 2016 [7] and 2017 [5] challenge datasets — including training, validation, and testing images along with their corresponding segmentation masks. The combined dataset is created by merging all images and masks from both years, followed by a deduplication step to remove overlapping samples. This is necessary because a portion of the ISIC 2017 dataset reuses samples from ISIC 2016.

We refer to the final merged and deduplicated dataset as ISIC-Merged. The International Skin Imaging Collaboration (ISIC) is one of the largest archives of skin lesion images, largely due to its aggregation of data from various sources. While the vast amount of data benefits model training, the lack of standardization introduces challenges. It is in fact important to note that the images exhibit significant variability: some have vertical borders, others have circular borders, and a few contain calibration stickers. Figure 2 illustrates such inconsistency within the data. The absence of uniformity among these data sources has the potential to yield mediocre results when employed in the training of neural networks.

We perform a 60-20-20 split on ISIC-Merged to obtain the training, validation and testing sets.

3.1.1 Preprocessing

The images from the ISIC-Merged dataset are first processed to remove hair using a black-top-hat filter, as described in Ashraf *et al*. [1]. Subsequently, the images are resized to 256x256 pixels. An example of preprocessing is shown in Figure 1.



(a) Original image (b) Preprocessed image

Figure 1: Comparison of images (a) before and (b) after application of top black hat filter and resizing

3.1.2 Data augmentation

We employ multiple independent data augmentation techniques, resulting in a training dataset expansion proportional to the number of augmentation methods applied. While conventional augmentation approaches such as rotation and horizontal flipping are utilized, the principal contribution of our work is a novel augmentation strategy specifically designed to address the critical underrepresentation of dark skin tones in skin lesion datasets.

Our approach leverages ground truth segmentation masks as precise boundary identifiers to differentiate between lesion and surrounding skin regions. This differentiation enables targeted, region-specific modifications rather than applying uniform transformations across entire images. The resulting region-aware augmentation offers significantly more nuanced and clinically relevant transformations compared to traditional whole-image augmentation techniques.

The proposed methodology comprises two distinct variations designed to simulate different clinical presentations in diverse skin tones:

1. **Dark Skin Augmentation 1 (DS1)**: This variation maintains higher brightness in the lesion re-

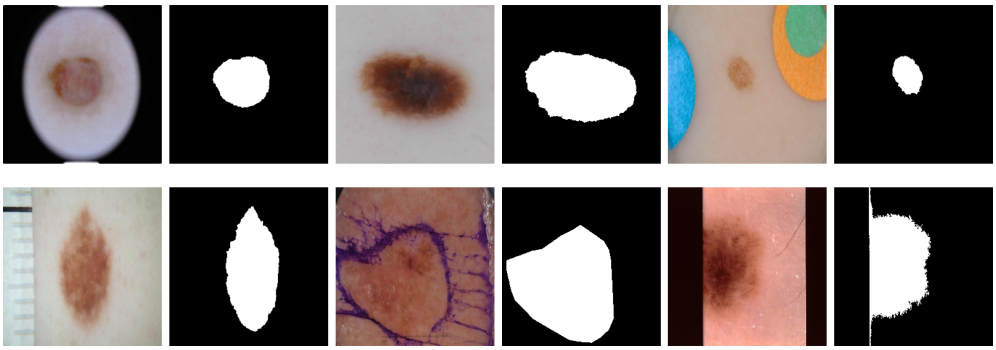


Figure 2: Samples extracted from the preprocessed ISIC-Merged with the respective segmentation mask.

gion while darkening the surrounding skin. This configuration accurately simulates hypopigmentation characteristics frequently observed in lesions on darker skin tones, where affected areas present as lighter than surrounding tissue due to melanin disruption (see Figure 3).

2. **Dark Skin Augmentation 2 (DS2)**: This variation significantly darkens the lesion region while modifying the surrounding skin characteristics. The deliberate reduction in contrast between lesion and surrounding tissue challenges the neural network with low-contrast scenarios, thereby enhancing detection capabilities in situations where lesion visibility is compromised—a common challenge in clinical dermatology, particularly with darker skin tones (see Figure 4).

In both DS1 and DS2 variations, we implement a comprehensive suite of targeted modifications exclusively to the surrounding skin region (non-lesion areas), preserving the integrity of the lesion characteristics while transforming the surrounding tissue. These modifications include: precise adjustments to brightness and saturation levels, calibrated alterations of red and green color curves, HSL (Hue, Saturation, Lightness) transformations, controlled application of Gaussian blurring, and strategic introduction of noise. It is important to note that the HSL transformations and color curve adjustments are applied exclusively to the outer region (normal skin) and not to the lesion itself, maintaining the critical diagnostic features of the lesion while altering the surrounding context.

This sophisticated combination of region-specific techniques enables the simulation of diverse skin tones, with particular emphasis on darker complexions that remain significantly underrepresented in standard dermatological datasets. The resulting augmented images preserve diagnostically relevant features while presenting them in contexts that more accurately reflect the diversity of patient populations.

Our investigation takes into consideration the following experimental configurations:

Configuration	Dataset Composition
Case 1	ISIC-Merged
Case 2	ISIC-Merged + horizontal flip + 90° rotation + hue shift + Gaussian noise
Case 3	ISIC-Merged + horizontal flip + 90° rotation + hue shift + Gaussian noise + stretching
Case 4	ISIC-Merged + horizontal flip + 90° rotation + hue shift + Gaussian noise + stretching + DS1 + DS2

Table 2: Experimental configurations

3.1.3 Dark skin test samples

Our evaluation of the skin cancer detection model's performance on dark skin tones is constrained by a critically limited dataset of only 8 images from the University of Waterloo Skin Cancer Detection Dataset [16]. While this small sample size is insufficient for comprehensive validation, it provides a preliminary indication of potential improvements in segmentation accuracy (Figure 8).

3.2. Network Architecture

We begin by implementing a basic UNet architecture, a fully convolutional network originally designed for biomedical image segmentation by Ronneberger *et al*. [17]. The architecture, illustrated in Figure 5b, follows an encoder–decoder structure: the encoder progressively downsamples the input to extract high-level

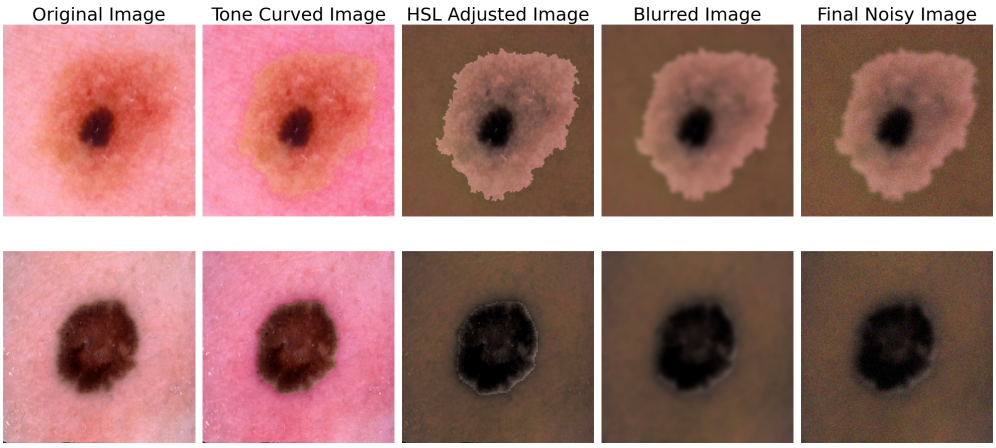


Figure 3: Modifications in the DS1 Data Augmentation Technique

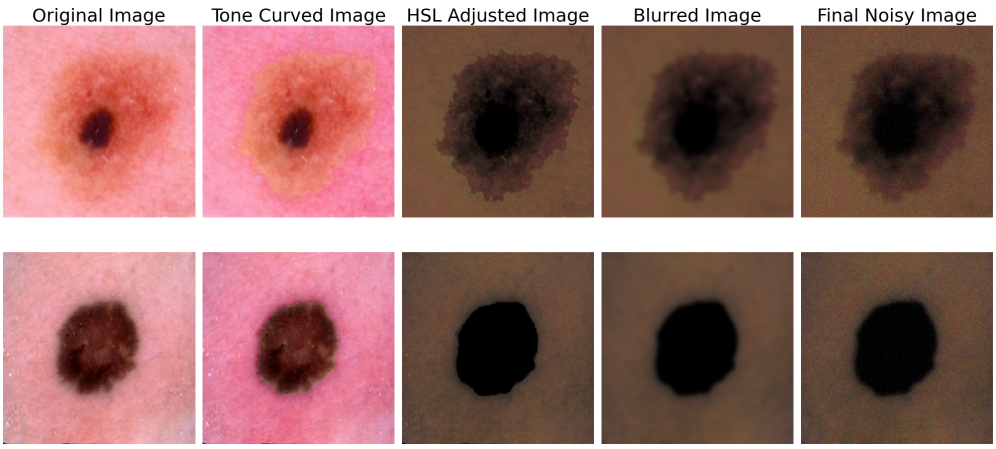


Figure 4: Modifications in the DS2 Data Augmentation Technique

features, while the decoder upsamples to reconstruct the segmentation map. Skip connections between corresponding encoder and decoder layers help preserve spatial information and improve segmentation accuracy.

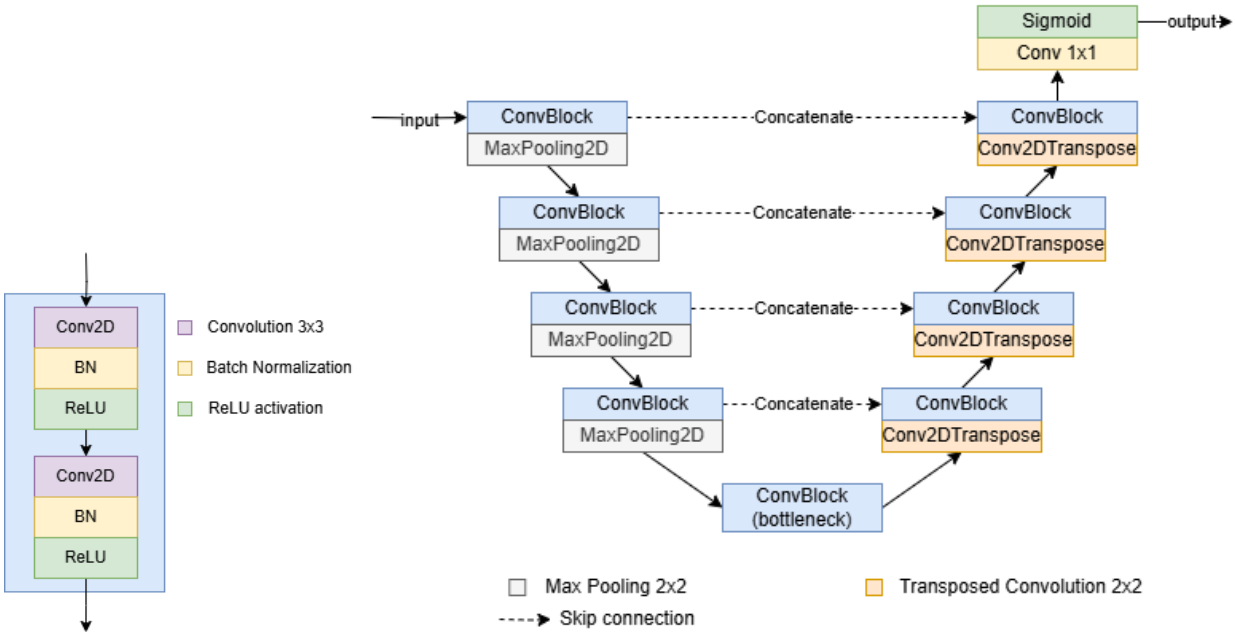
After experimenting with different depths, we selected a depth of 4 which is the point where we start to get only marginal gains in validation performance on ISIC-Merged, as shown in Figure 7.

We then extended the architecture by incorporating attention mechanisms—specifically SCSE and cSE modules [18]—as illustrated in Figure 6. The channel Squeeze-and-Excitation (cSE) module applies channel-wise attention, selectively emphasizing informative feature channels while suppressing less relevant ones. The Spatial and Channel Squeeze-and-Excitation (SCSE) module extends this approach by incorporating both spatial and channel-wise recalibration, allowing for more comprehensive feature refinement. These mechanisms are hypothesized to improve the network discriminative capability by accentuating anatomically relevant structures while attenuating irrelevant ones, thereby potentially increasing segmentation accuracy.

3.3. Training

The network is trained end-to-end using the ADAM [10] optimizer with a hybrid loss function composed of 10% Binary Cross-Entropy (BCE) and 90% Dice Loss [20]. A higher weight is assigned to Dice Loss, as it better captures object boundaries, addresses class imbalance, and improves segmentation performance, particularly for small or irregularly shaped objects [13].

The Binary Cross-Entropy (BCE) loss is defined as:



(a) ConvBlock (b) Basic UNet

Figure 5: Basic UNet [17] architecture of depth 4.

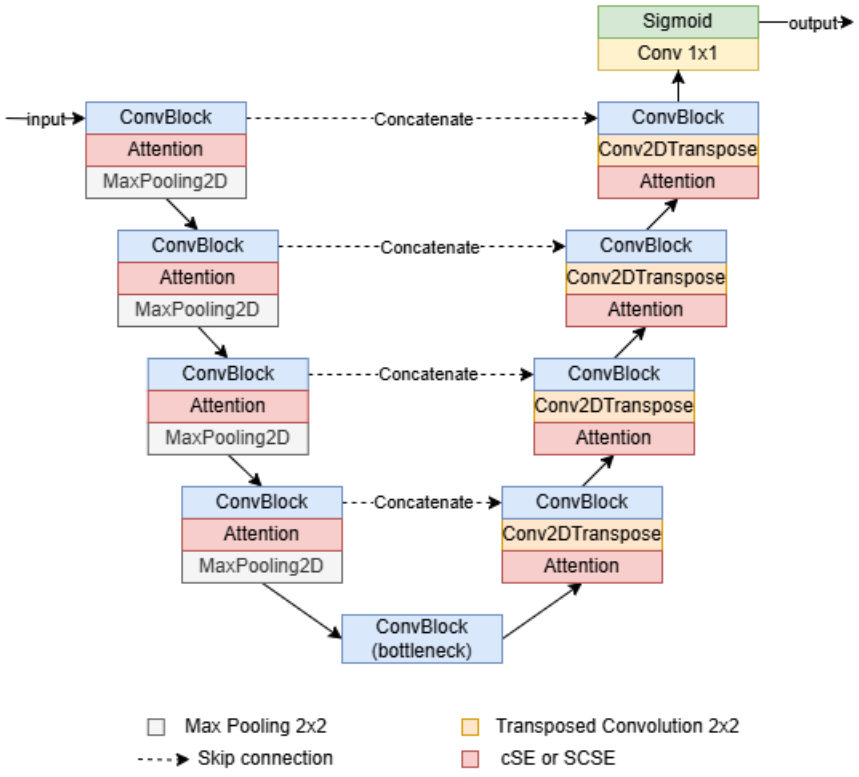


Figure 6: UNet extended with cSE or SCSE [18] attention mechanisms.

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right] \tag{1}$$

where y_i is the ground truth label, $\hat{y}_i$ is the predicted probability, and N is the total number of pixels.

The Dice Loss is formulated as:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N} y_i \hat{y}_i + \epsilon}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y}_i + \epsilon} \tag{2}$$

where ϵ is a small constant to prevent division by zero. We use $\epsilon = 1 \times 10^{-6}$.

The model is trained for a maximum of 200 epochs, with early stopping triggered if the validation loss on the 20% ISIC-Merged validation set does not decrease for 16 consecutive epochs. The initial learning rate is set to 1×10^{-3} and is reduced by half if validation loss stagnates for 5 epochs, down to a minimum of 1×10^{-6}. At the end of training, the model weights corresponding to the lowest validation loss are restored.

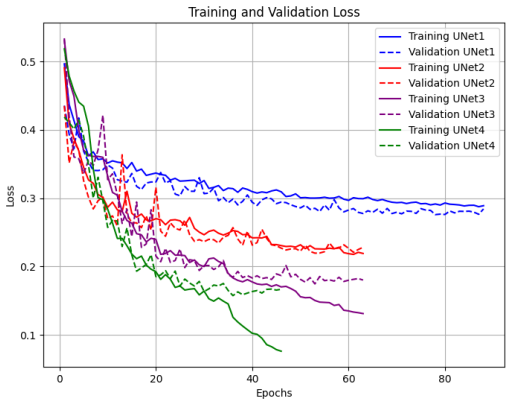


Figure 7: Training and validation loss history on ISIC-Merged split for increasingly deep UNets.

4. Results

We evaluate the performance of our models using the Jaccard Index, also known as the Intersection over Union (IoU), a widely-used metric in semantic segmentation tasks. The Jaccard Index quantifies the overlap between the predicted segmentation $\hat{Y}$ and the ground truth mask Y, and is defined as:

$$\text{Jaccard Index} = \frac{|Y \cap \hat{Y}|}{|Y \cup \hat{Y}|} = \frac{TP}{TP + FP + FN} \tag{3}$$

where TP refers to the number of true positive pixels, FP to false positives, and FN to false negatives. A higher Jaccard Index indicates better segmentation performance.

Model Configuration	Case 1	Case 2	Case 3	Case 4
Basic UNet	0.7485	0.7561	0.7576	0.7698
cSE UNet	0.7413	0.7474	0.7732	**0.7832**
SCSE UNet	0.7180	0.7498	0.6938	0.7016

Table 3: Jaccard Index scores achieved across different architectures and augmentation setups.

4.1. Analysis

Table 3 reports the Jaccard Index across different augmentation setups and UNet architectural variants. The baseline UNet trained only on the original ISIC-Merged dataset (Case 1) achieves a score of 0.7485. Its performance gradually improves with each additional augmentation stage, reaching 0.7698 in Case 4. This trend suggests a clear benefit from the use of data augmentation.

The UNet with SCSE attention exhibits inconsistent performance across cases. Its performance in Case 1 is worse than the baseline. While it improves from 0.7180 in Case 1 to 0.7498 in Case 2, its score drops significantly in Case 3 (0.6938) and only marginally recovers in Case 4 (0.7016). These results suggest that the SCSE mechanism may be sensitive to certain transformations, such as stretching and synthetic dark-skin augmentation.

In contrast, the cSE-enhanced UNet performs more robustly. It starts slightly below the baseline in Case 1 (0.7413) but steadily outperforms the other models in the later configurations. It achieves 0.7474 in Case 2, 0.7732 in Case 3, and reaches the highest overall score of 0.7832 in Case 4.

4.2. Discussion

The observed trends highlight the complementary roles of data augmentation and attention mechanisms in segmentation performance. The basic UNet consistently benefits from more diverse and complex data, confirming that progressive augmentation increases the network ability to generalize.

SCSE attention appears to be destabilized by the more aggressive augmentations, particularly those altering spatial characteristics. The drop in performance from Case 2 to Case 3 indicates that spatial recalibration mechanisms may be sensitive to geometric transformations that distort image structure such as stretching.

On the other hand, the cSE module—focused solely on channel-wise feature recalibration—exhibits resilience across all settings. It benefits most from

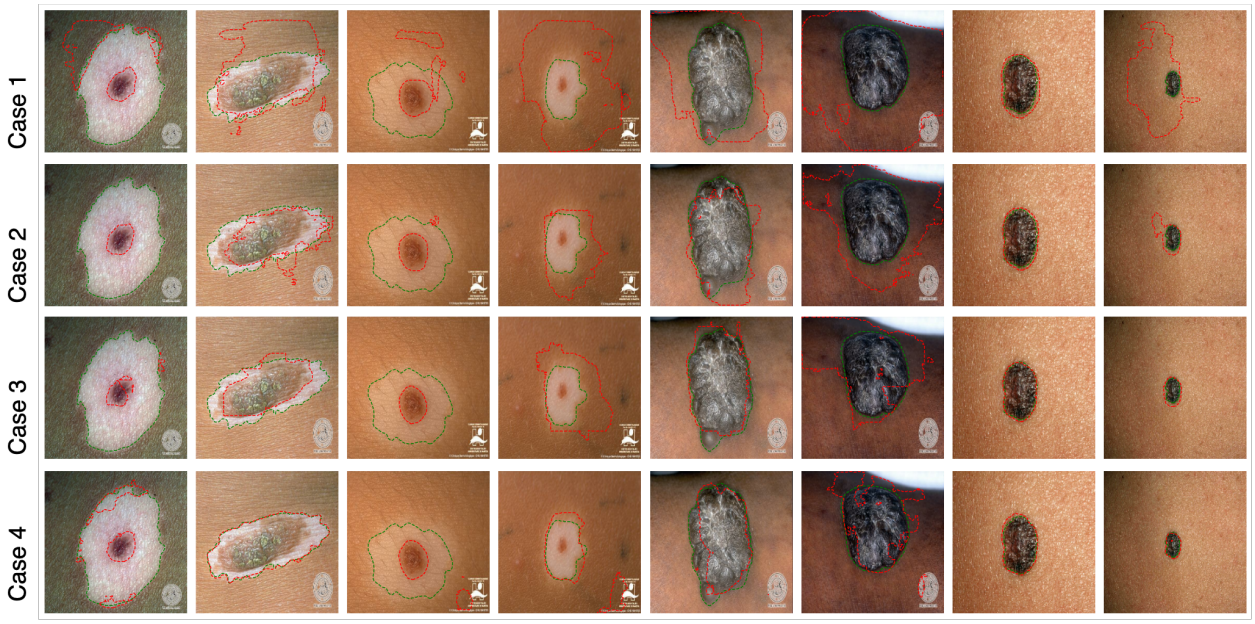


Figure 8: Segmentation performance on dark skin samples achieved by the basic UNet models trained on the configurations described in Table 2. The green line is the ground truth and the red one is the prediction.

richer, region-aware augmentations introduced in Cases 3 and 4. These results suggest that channel attention is better suited for segmentation tasks involving substantial dataset diversity, especially when spatial consistency is less predictable.

These findings emphasize that attention mechanisms must be aligned with both task requirements and data augmentation characteristics. While augmentations can improve robustness, the architecture must be capable of leveraging such diversity effectively.

Figure 8 reveals nuanced improvements for the 8 dark skin samples 3.1.3 through the employed augmentation techniques. For samples 2 and 4, the neural network demonstrates a significant advancement by successfully identifying both the dark and clear portions of the lesion—a critical capability absent in the initial model. Samples 6 and 8, which previously presented significant segmentation difficulties due to poor lighting, now exhibit markedly improved accuracy. Conversely, samples 1 and 3 display more limited progress, with the network continuing to segment only the darker lesion portions, indicating that the augmentation techniques' effectiveness varies across different image characteristics.

5. Conclusion

This work explores data augmentation strategies and attention-based architectural modifications to improve skin lesion segmentation, with a particular focus on addressing the underrepresentation of dark-skinned patients. We introduce region-aware augmentations (DS1 and DS2) that simulate skin tone variations using mask-guided transformations. Our findings show that these augmentations enhance overall segmentation performance, especially when combined with channel-based attention.

Among all configurations, the UNet with cSE blocks trained on the most comprehensive augmented dataset achieves the highest Jaccard Index of 0.7832. This result highlights the potential of channel-wise recalibration in handling diverse input data while maintaining segmentation accuracy.

However, important limitations remain. The ISIC datasets suffer from a lack of standardization, with significant variability in image quality, acquisition conditions, and skin tones. This inconsistency can impact the effectiveness of augmentations and lead to biased model evaluations. Furthermore, while DS1 and DS2 augmentations aim to improve fairness, their effectiveness is constrained by the heterogeneity of the base dataset.

6. Future Work

Several avenues for future research remain critical. The primary limitations include dataset diversity, with our current research being constrained by a limited number of dark-skinned images on which to test the proposed methods, suggesting the need for more comprehensive and representative datasets. Real-world clinical validation remains essential, requiring rigorous test-

ing on diverse clinical datasets and expert dermatological evaluation to bridge the gap between algorithmic advancements and practical medical implementation, ultimately ensuring diagnostic accuracy and algorithmic fairness across diverse demographic groups.

Future work should prioritize the construction of diverse but standardized dermatological image datasets, accompanied by benchmarks focused on equity and diagnostic reliability. Until such datasets are available, region-aware augmentation provides a practical approach to addressing skin tone imbalance in existing datasets, offering a foundation for more inclusive and fair dermatological AI systems. Future work could also refine the DS1 and DS2 augmentation techniques that, due to the previously mentioned diversity and lack of standardization of the data, in some cases lead to excessively dark or low-saturated images. To address this limitation, one of the next steps might be classifying images based on skin color as previously done by Benčević et al. [2] and applying different transformations based on the skin color and lighting conditions of the image.

References

[1] Hassan Ashraf, Asim Waris, Muhammad Fazeel Ghafoor, Syed Omer Gilani, and Imran Khan Niazi. Melanoma segmentation using deep learning with test-time augmentations and conditional random fields. *Scientific Reports*, 12(1):3948, 2022. 1, 2

[2] Marin Benčević, Marija Habijan, Irena Galić, Danilo Babin, and Aleksandra Pižurica. Understanding skin color bias in deep learning-based skin lesion segmentation. *Computer methods and programs in biomedicine*, 245:108044, 2024. 1, 8

[3] Alceu Bissoto, Michel Fornaciali, Eduardo Valle, and Sandra Avila. (de) constructing bias on skin lesion datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1

[4] Alceu Bissoto, Eduardo Valle, and Sandra Avila. Gan-based data augmentation and anonymization for skin-lesion analysis: A critical review. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1847–1856, 2021. 1, 2

[5] Noel C Codella, David Gutman, M Emre Celebi, Blake Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan C Halpern. Skin lesion analysis toward melanoma detection 2017: A challenge hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 168–172. IEEE, 2018. 2

[6] Roxana Daneshjou, Kailas Vodrahalli, Roberto A Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M Swetter, Elizabeth E Bailey, Olivier Gevaert, et al. Disparities in dermatology ai performance on a diverse, curated clinical image set. *Science advances*, 8(31):eabq6147, 2022. 1

[7] David Gutman, Noel Codella, M Emre Celebi, Blake Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1605.01397*, 2016. 2

[8] Gery P Guy Jr, Cheryll C Thomas, Trevor Thompson, Meg Watson, Greta M Massetti, Lisa C Richardson, Centers for Disease Control, Prevention (CDC), et al. Vital signs: melanoma incidence and mortality trends and projections-united states, 1982-2030. *MMWR Morb Mortal Wkly Rep*, 64(21):591–596, 2015. 1

[9] Anthony F Jerant, Jennifer T Johnson, Catherine Demastes Sheridan, and Timothy J Caffrey. Early detection and treatment of skin cancer. *American family physician*, 62(2):357–368, 2000. 1

[10] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. https://arxiv.org/abs/1412.6980, 2017. 4

[11] Ammara Masood and Adel Ali Al-Jumaily. Computer aided diagnostic support system for skin cancer: a review of techniques and algorithms. *International journal of biomedical imaging*, 2013(1):323268, 2013. 1

[12] Agnieszka Mikołajczyk, Sylwia Majchrowska, and Sandra Carrasco Limeros. The (de) biasing effect of gan-based augmentation methods on skin lesion images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 437–447. Springer, 2022. 1, 2

[13] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016. 4

[14] Zahra Mirikharaji, Kumar Abhishek, Alceu Bissoto, Catarina Barata, Sandra Avila, Eduardo Valle, M Emre Celebi, and Ghassan Hamarneh. A survey on deep learning for skin lesion segmentation. *Medical Image Analysis*, 88:102863, 2023. 1

[15] Andres Morales-Forero, Lili Rueda Jaime, Sebastian Ramiro Gil-Quiñones, Marlon Y Barrera Montañez, Samuel Bassetto, and Eric Coatanea. An insight into racial bias in dermoscopy repositories: A ham10000 data set analysis. *JEADV Clinical Practice*, 3(3):836–843, 2024. 1

[16] University of Waterloo. Skin cancer detection (2016). https://uwaterloo.ca/vision-image-processing-lab/research-demos/skin-cancer-detection. 3

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. https://arxiv.org/abs/1505.04597, 2015. 3, 5

[18] Abhijit Guha Roy, Nassir Navab, and Christian Wachinger. Recalibrating fully convolutional networks with spatial and channel 'squeeze excitation' blocks. https://arxiv.org/abs/1808.08127, 2018. 4, 5

[19] Robert S Stern. Prevalence of a history of skin cancer in 2007: results of an incidence-based model. *Archives of dermatology*, 146(3):279–282, 2010. 1

[20] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 240–248. Springer, 2017. 4